Extensive Population Structure in San, Khoe, and Mixed Ancestry Populations from Southern Africa Revealed by 44 Short 5-SNP Haplotypes

Article *in* Human Biology · December 2012

DOI: 10.3378/027.084.0603 · Source: PubMed

1-1-2012

Extensive Population Structure in San, Khoe and Mixed Ancestry Populations from Southern Africa Revealed by 44 Short 5-SNP Haplotypes

Carina M. Schlebusch
Human Genomic Diversity and Disease Research Unit, Division of Human Genetics, School of Pathology, Faculty of Health Sciences, University of the Witwatersrand and National Health Laboratory Service. Johannesburg, 2000, South Africa, cschlebu@gmail.com

Himla Soodyall
Human Genomic Diversity and Disease Research Unit, Division of Human Genetics, School of Pathology, Faculty of Health Sciences, University of the Witwatersrand and National Health Laboratory Service. Johannesburg, 2000, South Africa

Extensive population structure in San, Khoe and Mixed Ancestry populations from southern Africa revealed by 44 short 5-SNP haplotypes

Carina M Schlebusch[1,2§]*, Himla Soodyall* [1]

[1] Human Genomic Diversity and Disease Research Unit, Division of Human Genetics, School of Pathology, Faculty of Health Sciences, University of the Witwatersrand and National Health Laboratory Service. Johannesburg, 2000, South Africa

[2] Department of Evolutionary Biology, Evolutionary Biology Centre, Uppsala University, Norbyvägen 18D, SE-752 36 Uppsala, Sweden

Running title: Genotype and haplotype variation in populations from southern Africa

Keywords: Population structure, Haplotypes, Khoe-San, Khoisan-speaking, Coloured, southern Africa

[§]Corresponding author: Carina M. Schlebusch

Department of Evolutionary Biology, Evolutionary Biology Centre, EBC

Uppsala University, Norbyvägen 18D

SE-752 36 Uppsala, Sweden

Mobile number: +46 76 306 3341

Fax: +46 18 4716310

Email: cschlebu@gmail.com

ABSTRACT

The San and Khoe people currently represent remnant groups of a much larger and widely distributed population of hunter gatherers and pastoralists who had exclusive occupation of southern Africa before the arrival of Bantu-speaking groups in the past 1,200 years and sea-borne immigrants within the last 350 years. Genetic studies (mitochondrial DNA and Y-chromosome) conducted on San and Khoe groups revealed that they harbour some the most divergent lineages found in living peoples throughout the world. Recently, high-density autosomal SNP-array studies confirmed the early divergence of Khoe-San population groups from all other human populations. The present study made use of 220 autosomal SNP markers, in the format of both haplotypes and genotypes, to examine the population structure of various San and Khoe groups and their relatedness to other neighbouring groups.

While analyses based on the genotypic SNP data only supported the division of the included populations into three main groups, Khoe-San, Bantu-speakers and non-African populations, haplotype analyses revealed finer structure within Khoe-San populations. Through using only 44 short SNP haplotypes (compiled from a total of 220 SNPs), most of the Khoe-San groups could be resolved as separate groups by applying STRUCTURE analyses. Therefore, by carefully selecting a few SNPs and combining them into haplotypes, we were able to achieve the same level of population distinction as achieved previously in high-density SNP studies on the same population groups. Using haplotypes proved to be a very efficient and cost-effective way to study population structure.

INTRODUCTION

Africa has remarkable cultural, linguistic and genetic diversity and more than 2,000 distinct ethnic groups and languages exist on the continent (Gordon, 2005). All genetic studies to date provide substantial support for a predominantly African origin of modern humans. The greatest genetic variation has consistently been identified within African populations and variation outside of Africa has been shown to be a subset of the African diversity (Garrigan and Hammer, 2006, Jobling and Tyler-Smith, 2003, Torroni et al., 2006, Underhill and Kivisild, 2007, Conrad et al., 2006).

The majority of sub-Saharan Africans (>200 million people) speak one of ~500 very closely related languages, these languages are collectively referred to as "Bantu" languages, based on the word meaning "people" (Bleek, 1862). The current distribution of these groups is largely a consequence of the movement of people (demic diffusion) rather than a diffusion of only language (Ehret and Posnansky, 1982, Huffman, 1982). This expansion, commonly referred to as the "Bantu Expansion" (Greenberg, 1963), began ~3,000 – 5,000 years BP (Ehret and Posnansky, 1982, Vansina, 1990) and is thought to be due to the development and spread of agriculture and, possibly, the use of iron (Greenberg, 1972, Newman, 1995, Phillipson, 1993). To a certain extent, the expansions of Bantu-speaking groups masked the earlier history of non-Bantu-speaking African populations.

Groups that existed all over the African continent before the Bantu-expansions were replaced and/or assimilated by the Bantu-speaking groups. Traces of these pre-Bantu groups might still be found in the genetic variation, language and cultural practices of various Bantu-speaking groups where they have been incorporated or assimilated. However, few sub-Saharan African ethnic groups have retained a cultural, linguistic and genetic identity that distinguishes them from the Bantu-

speaking groups. Examples of such groups of people are the Hadza and Sandawe from East Africa, the Khoe-San populations from southern Africa and the Pygmy populations from central Africa. These populations (excluding the Khoe) did not adopt an agricultural lifestyle but instead kept a hunter-gatherer lifestyle. Their cultural practices, lifestyle and language (for the Khoe-San, Hadza and Sandawe) distinguished them from Bantu-speakers. An increasing number of these groups are now adopting or have recently adopted food producing and/or sedentary lifestyles. Their distinction from other groups, however, is still visible in the comparative genetic analysis of these populations in relation to Bantu-speakers and other neighbouring populations. In both Y-chromosome and mitochondrial DNA studies, these hunter-gatherer populations tend to carry unique and older lineages than the lineages associated with the Bantu-speaking people. Some of the most divergent haplogroups known among modern humans, for mitochondrial DNA and the Y-chromosome, are found commonly and at their highest frequencies in the Khoe-San people (Behar et al., 2008, Chen et al., 2000, Karafet et al., 2008, Knight et al., 2003, Naidoo et al., 2010, Schlebusch et al., 2009, Scozzari et al., 1999, Tishkoff et al., 2007, Underhill et al., 2001, Batini et al., 2011, Schlebusch et al., 2011). Additionally, in autosomal studies, San people group in a distinct cluster from that of Bantu-speakers (Cavalli-Sforza et al., 1994, Tishkoff et al., 2009, Li et al., 2008, Jakobsson et al., 2008, Rosenberg et al., 2002, Schlebusch et al., 2012, Pickrell et al., 2012). Thus, these unique populations of hunter-gatherers who carry genetic variation belonging to the deepest clades known among modern humans are crucial links to the past. However, it is becoming increasingly difficult to study these groups, as the Khoe-San groups are losing their cultural identities, lifestyles and languages and are integrating into surrounding groups.

The term “Khoe-San” has a collective meaning for two groups of people, the Khoi (old Nama word) or Khoe (modern Nama word), who were traditionally the pastoralist groups and the San, which included the hunter-gatherer groups (Schlebusch, 2010, Crawhall, 2006). This grouping was made according to the traditional division that existed between hunter-gatherers and pastoralists. Different San and Khoe groups are distributed throughout southern Africa where they live among and to some extent are admixed with the various Bantu-speaking populations surrounding them (Barnard, 1992, Smith et al., 2000, le Roux and White, 2004). To classify Khoe-San groups into their individual ethnic groups is, in many ways, problematic. Different words and spellings have been used to refer to the same groups of people over the years. Linguistic classification is the method most commonly used to identify different groups, but it is not clear that linguistic classification reflect genetic relationships.

Linguistic studies indicate three separate linguistic families for the southern Khoisan linguistic division and East Africa have two additional languages classifying under Khoisan namely Hadza and Sandawe. The three southern Khoisan language groups are; Ju (previously classified as Northern Khoisan), Tuu (previously classified as Southern Khoisan) and Khoe (previously classified as Central Khoisan) (Güldemann, 2008). These linguistic families are either unrelated or have genealogical relationships that extend further back than 10,000 years (Güldemann, In Press). Linguistic evidence support the possibility that the Ju and Tuu branches may share a very deep common ancestor and were associated with the original San hunter-gathers, while the Khoe branch was introduced to the area later in conjunction with pastoralism by an East African group (Güldemann, In Press, Güldemann, 2008). Accordingly a putative linguistic link between the Khoe linguistic branch and the East

African Sandawe Khoisan linguistic branch was proposed (Güldemann, In Press, Güldemann, 2008). From genetic studies, based on Y-chromosome markers in the !Xun and Khwe groups and various east African groups, it was theorized that the Khwe is a descendant group from the east African pastoralists that introduced sheep into southern Africa (Henn et al., 2008).

Although the San and Khoe groups are relatively small populations today, their genetic contribution to the Coloured population of South Africa may be substantial (de Wit et al., 2010, Quintana-Murci et al., 2010, Patterson et al., 2010). To understand the underlying genetic factors in an admixed population, it is important to study the parental populations.

Various genetic studies published results on "Khoisan" groups (Chen et al., 2000, Tishkoff et al., 2007, Vigilant et al., 1991, Scozzari et al., 1997, Cruciani et al., 2002, Cruciani et al., 2004, Knight et al., 2003, Jobling and Tyler-Smith, 2003, Scozzari et al., 1999, Henn et al., 2008, Underhill et al., 2000, Underhill et al., 2001, Semino et al., 2002, Rosenberg et al., 2005, Rosenberg et al., 2002, Li et al., 2008, Jakobsson et al., 2008, Tishkoff et al., 2009). These studies, however, were not representative of all the linguistic families and collectively only included two Ju-speaking groups, namely, the Ju/'hoansi and the !Xun and one Khoe-speaking group, namely, the Khwe. Recently two independent studies described genetic results obtained from high-density genome-wide SNP-arrays, for an extensive collection of Khoe and San groups, representing all three main southern Khoisan linguistic divisions (Schlebusch et al., 2012, Pickrell et al., 2012). The current article present genetic results for 181 individuals that overlap with Khoe, San, Coloured and Bantu-speaking individuals from Schlebusch et al., (2012), an additional 102 individuals from these already mentioned population groups and 69 individuals from four other

population groups, namely; the South African Indian population (25 individuals), the South African Afrikaner population (15 individuals), a Bantu-speaking group from the DRC (14 individuals) and a group of random European origins (15 individuals) (Table 1). Results from the present study differs from the results presented in Schlebusch et al., (2012) in that a specific selection of 220 genome-wide SNPs (of which only 56 overlap with SNPs typed in Schlebusch et al., (2012)) are employed to effectively reveal population structure through the use of short inferred haplotypes.

Many studies have discussed the utility of employing haplotypes (Gattepaille and Jakobsson, 2012, Lawson et al., 2012, Morin et al., 2009, Browning and Weir, 2010) and it has been shown that combining closely situated markers (that are in Linkage Disequilibrium - LD), significantly improves the ability to assign individuals to population groups (Gattepaille and Jakobsson, 2012). Haplotype loci are multi-allelic and therefore more information about ancestry is available per locus. A further very useful and beneficial property of haplotype loci is that they are less affected by ascertainment bias (Browning and Weir, 2010). The available information on variant positions in the human genome (such as SNPs) has been predominantly obtained from certain population groups from specific geographic regions. This led to a situation where known SNP variants are biased towards SNPs at high frequency in European, East Asian and to a certain extent West African populations. SNP studies that do not correct for this inherent ascertainment bias will always overestimate genetic variation in these populations and underestimate genetic variation in other populations. Combing SNPs into haplotypes, greatly alleviate the effect of ascertainment bias because haplotypes are represented by multiple patterns of SNPs. In the present study we opted to make use of the advantageous properties of haplotype loci and therefore carefully selected a small number of SNPs in a very specific way to facilitate both

haplotypic and genotypic analyses. We then compared results from the haplotype analyses and the genotype analyses and discuss their relative abilities to uncover population sub-structure in our sample set.

MATERIALS AND METHODS

Subjects

DNA samples from 352 unrelated individuals were collected with the subjects' informed consent, and the project was approved by the Human Research Ethics Committee (Medical) at the University of the Witwatersrand, Johannesburg (Protocol Number: M050902), the Working Group of Indigenous Minorities in Southern Africa (WIMSA) and the South African San Council.

A description of sample groups, group codes, group membership, linguistic grouping, number of individuals, place of sampling and origin are outlined in Table 1. The overlap between samples of the present study and those of Schlebusch et al., (2012) is also summarised in Table 1.

Terms to describe the populations in this manuscript were chosen to be as unambiguous as possible while simultaneously being non-offensive to any population group. The people of mixed ancestry who participated in our study preferred to be classified as Coloured and did not perceive the term as derogatory. Furthermore, the term San and Khoe was used in the manuscript to refer to groups of people who are known descendants of the hunter-gatherers (San) and pastoralists (Khoe) that occupied southern Africa before Bantu-speaking groups arrived. San or Khoe and if need be Khoe-San, is the preferred term recommended by San communities (represented by the "Working Group of Indigenous Minorities in Southern Africa" and the "South African San Institute") (Schlebusch, 2010, Crawhall, 2006).

DNA extraction

DNA from EDTA-blood was extracted using the salting-out method described by Miller et al., (Miller et al., 1988) and the PureGene® Genomic DNA Purification Kit (Gentra Systems) was used to extract DNA from buccal swabs according to the manufacturer's instructions.

SNP selection and typing

A total of 220 autosomal SNPs were specifically selected in the following way: 10 SNPs per chromosome (chromosome 1 to 22) were selected; the 10 SNPs per chromosome were selected in two groups of 5 linked SNPs; the two groups of 5 linked SNPs were completely unlinked from one another (Figure A1, Appendix). The five SNPs in the 5 SNP group were selected to be on the same haploblock. To select SNPs on the same haploblock the software SNPbrowserTM v3.1 (Applied Biosystems) were used and both HapMap and Applied Biosystems (ABI) SNP databases were considered. In the ABI database, haplotype blocks from the African American study group were considered and in the HapMap database, haplotype blocks from the Yoruba study group were considered. None of these two study groups are Khoe-San but these were the closest related population groups from which sufficient SNP data was available at the time of SNP selection. SNPs were selected to be on the same haploblock in the Yoruba group and preferentially also on the same haploblock in the African American group. The average distance between consecutive selected SNPs in the same haploblock was 4347 bp (STD =3730.8 bp) (Table A2, Appendix). The haploblocks that contained the SNPs were not associated with any known coding part of the genome, therefore neutral genetic variation was targeted and

influence of selection minimized. Furthermore SNPs were selected to have a minor allele frequency above 10% in the African population groups. The full list and details of selected SNPs are included in Table A2 in the Appendix.

SNPs were selected in this fashion to allow for multiple types of analyses using the same dataset. Firstly the selection allows for the compilation of multiple different genotype sets with 44 unlinked polymorphisms in each, by selecting one SNP per SNP-group (of 5 linked SNPs) (see Figure A1, Appendix). Furthermore haplotypes can be inferred for SNPs on the same haploblock, these short inferred haplotypes can be used in similar and additional analyses than for unlinked genotypic SNPs (see below) and results can be compared. In the haplotype based analyses each 5-SNP set were treated as a multi-allelic locus where each allele is represented by a unique 5-SNP haplotype.

All autosomal SNPs were typed commercially (by Harvard-Partners Centre for Genetics and Genomics, Genotyping Facility, Cambridge, Massachusetts, United States), by means of Sequenom iPLEX SNP genotyping. Multiplex PCR assays were designed using Sequenom SpectroDESIGNER software (version 3.0.0.3). Genotypes generated were subjected to QC and seven of the 220 loci were excluded because of poor assay quality (indicated Table A2 in the Appendix).

Data analyses

For genotypic analysis, 100 different genotype datasets with 44 unlinked SNP polymorphisms in each were compiled by randomly selecting one SNP per SNP-group (of 5 linked SNPs) from the same typed set of 220 SNPs (see Appendix Figure A1). To generate the haplotypic dataset, the five linked SNPs on the same haploblock was used to infer 44 haplotypes consisting of 5 bp each. The haplotypes were inferred

separately for each population and each SNP set of 5 using POWERMARKER v3.25 (Liu and Muse, 2005).

Expected heterozygosity (per population) for the 100 different genotypic SNP datasets as well as the haplotypic dataset, were calculated using POWERMARKER v3.25 (Liu and Muse, 2005). For the 100 genotypic datasets, the mean expected heterozygosity as well as the standard deviation across the 100 datasets were then calculated.

The mean number of distinct haplotype alleles per locus (allelic richness) and the mean number of private haplotype alleles per locus (private allelic richness) were computed for each population in the haplotypic dataset, using ADZE v.1.0 (Szpiech et al., 2008). The program uses a rarefaction method (Kalinowski, 2004) that correct for the sample size across populations.

Population structure signified by the short haplotypes and genotypes was assessed using the clustering approach implemented in STRUCTURE v2.2 (Pritchard et al., 2000, Falush et al., 2007, Falush et al., 2003). For the haplotypic dataset, we replicated the STRUCTURE analysis 10 times for each choice of assumed clusters (K), from K = 2 to K = 10. Each replicate STRUCTURE run used a burn-in period of 50,000 repeats, followed by 100,000 repeats. Allele frequencies were correlated and a model with admixture was assumed for all runs. The 10 replicates for each choice of K were summarized with CLUMPP version 1.1.1 (Jakobsson and Rosenberg, 2007) to identify common modes among replicates. The CLUMPP analysis used the LargeKGreedy algorithm with 10,000 random permutations. Common solutions were identified by looking at the CLUMPP pairwise G' values. All pairs with a symmetric similarity coefficient G' > 0.9 were selected to be representative of a single mode. For each K we used the most frequently occurring mode identified and ran CLUMPP a

second time (using the LargeKGreedy algorithm and 10,000 random permutations), using only the replicates belonging to this mode. From the second analysis, we obtained the mean across replicates of the cluster membership coefficients of each individual, for each mode at each value of K. The clustering results were visualized with DISTRUCT (Rosenberg, 2002).

For genotypic datasets, STRUCTURE was run on the 100 different SNP sets separately. The STRUCTURE analyses of the 100 sets of 44 unlinked SNPs were conducted with the same parameters as for the short haplotypes. For the summarizing of STRUCTURE output, the 10 iterations at each K for each of the 100 SNP sets were collapsed into one consensus run using CLUMPP (Jakobsson and Rosenberg, 2007) using the same procedure as described for haplotypes. Thereafter the consensus runs of the 100 sets of random SNPs were collapsed into one consensus run at each K using CLUMPP v1.1.1 and visualized using DISTRUCT (Rosenberg, 2002).

We furthermore constructed population distance matrices for both haplotypic and genotypic datasets. The same haplotypic dataset and 100 genotypic datasets used in the STRUCTURE analysis were also used in distance based analysis. To construct population distance matrices of the haplotypic dataset and each of the 100 genotypic datasets, Reynolds distance (Reynolds et al., 1983) was used as implemented in POWERMARKER v3.25 (Liu and Muse, 2005) and Neighbour Joining (NJ) trees were constructed from each matrix. To condense the 100 different NJ trees generated from genotypic data into one output, a Majority Rule consensus tree were constructed using CONSENCE implemented in PHYLIP v.3.65 (Felsenstein, 2004).

RESULTS

Summary statistics

Expected heterozygosities for populations based on the 44 short 5-SNP haplotypes differed from the mean expected heterozygosity estimated based on 100 random combinations of 44 unlinked SNPs (Table 2). For both data types, non-African populations had lower estimates compared to African populations. However, for estimates based on genotypic data, the Bantu-speaking and admixed populations generally had the highest estimates and Khoe and San groups lower, while for the haplotype based estimates, both Khoe-San and Bantu-speakers rank among the top and bottom African populations. The Khoe population, the Nama, had the highest estimates followed by the southeast Bantu-speakers and the Karretjie People, a group with San ancestry.

The allelic diversity of inferred haplotypes was further investigated by calculating the mean number of distinct alleles and mean number of private alleles as a function of a standardized sample size (Figure 1). The Nama, Karretjie People and Khwe were the richest in distinct alleles per locus, whereas the non-African populations had the fewest distinct alleles per locus. The mean number of the private alleles per locus was the highest for the /Gui + //Gana, the Nama and the Ju/'hoansi group, while the non-African groups and southern Bantu-speaking groups (SEB and HER) had the lowest frequency.

STRUCTURE analyses

The averaged results of the STRUCTURE runs for the 100 different genotypic SNP sets (of 44 unlinked SNPs) are shown in Figure 2A and B and Table A1 (Appendix). The iterations were done for K=2 to K=10. Iterations for K=2 to K=5 are shown (only K=2 and K=3 contained population structure information).

The highest resolution obtained with the genotypic data was three discernible groups (at K=3), namely non-African, Khoe-San and Bantu-speakers. Clustering at

K=2 separated non-African from African, at K=3 the African cluster is divided into a Khoe-San and Bantu-speaking component, while further clustering failed to distinguish more structure within the dataset (higher order clustering [K=4 to K=10] continued to resolve the Bantu-speaker cluster internally with no further substructure between study populations).

Different amounts of admixture into the different Khoe-San groups from Bantu-speakers and non-Africans were clearly visible at K=2 and K=3 (Appendix Table A1). The southern Khoe-San and Coloured groups (KAR, COL, CAC and also NAM) all had more input from the non-African cluster compared to the northern (JOH, XUN) and central San (KWE, GUG) and Bantu-speakers. The southern group with the least amount of admixture from non-African groups was the Karretjie People. The Karoo-Coloured group that resides nearby the Karretjie People had much higher contributions from the non-African cluster. The Cape-Coloured group had the highest input from the non-African cluster.

Except for the Ju/'hoansi group, STRUCTURE results supported asymmetric gene-flow between the Bantu-speakers and Khoe-San groups with more gene-flow from the Bantu-speakers into the Khoe-San than vice-versa. The Ju/'hoansi, !Xun, Karretjie People and Nama was the only groups where the Khoe-San cluster had a greater contribution than any of the other two clusters. In the /Gui + //Gana and Kgalagari group (GUG), the Bantu-speaking cluster contributed marginally more than the Khoe-San cluster. The Khwe had the largest input from the Bantu-speaking cluster of all the Khoe-San groups.

Gene-flow from the Khoe-San into the southern African Bantu-speakers (HER and SEB) is also evident. Seven present more input from the Khoe-San cluster was seen in the southern Bantu-speakers compared to the central African Bantu-speakers

(DRC) (Appendix Table A1). The Khoe-San component in the central African Bantu-speakers is most likely due to shared marker-ancestry (SNPs that are not variable between the two populations due to common ancestry before population divergence) and because of the low total number of SNPs in the genotypic dataset.

The European group also had some contribution from the African cluster (likely due to shared marker-ancestry and low numbers of typed SNPs), however, the increased African cluster allocation in the Afrikaner group compared to Europeans was probably due to recent admixture with African groups.

While inferences about admixture proportions could be made from the STRUCTURE analyses based on genotypes, finer levels of structure within Khoe-San groups were not evident. Haplotype analysis, however, revealed finer level structure to the extent that, at K=9, most Khoe-San groups had unique cluster components that identified their populations (Figure 2C and D). As was seen for the genotypic analysis K=2 and K=3 defined the non-African and Bantu-speaking components. However, contrasting with genotypic analysis, K=4 reveals sub-structure within the Khoe-San group with the two Ju-speaking northern San groups, !Xun and Ju/'hoansi, forming their own cluster separate from other Khoe-San groups. K=5, divides the Khoe-San into three clusters containing the 1) !Xun and Ju/'hoansi, 2) Khwe and /Gui, //Gana+Kgalagari and 3) Nama and Karretjie People. The two Coloured groups seem to attribute most of their Khoe-San component to these two newest clusters and not to the cluster representing the northern Ju-speakers. At K=6 a cluster emerges, which seems to have some representation in most of the groups, but has highest representation in the Khwe and when allowing seven clusters an additional Khwe cluster, uniquely associated with the Khwe, emerges (this cluster is mostly associated with specific Khwe individuals). With eight assumed clusters, the two northern Ju-

speakers, !Xun and Ju/'hoansi, split to form two separate clusters and at K=9, a cluster is associated with the Karretjie People, which separates them from the Nama group (although they still contain appreciable frequencies of the predominant Nama cluster, especially high in specific individuals).

Thus at K=9, we have obvious clusters associated with each of the following groups; non-Africans, Bantu-speakers, Ju/'hoansi, !Xun, /Gui, //Gana+Kgalagari, Nama and Karretjie People. The only Khoe-San group without a predominant specific cluster is the Khwe, which appears to be a highly mixed group with inputs from both Bantu-speakers and various Khoe-San groups, in addition to a small amount of unique genetic variation. In contrast to the substructure emerging for the Khoe-San groups, the Bantu-speaking and non-African clusters stayed homogenous, in spite of a large geographic divide between some of these groups.

Distance based analysis

The genotypic datasets (100 different genotypic SNP sets of 44 unlinked SNPs) and haplotype dataset (44 short 5-SNP haplotypes) were also used in distance based analysis, visualised in the form of radial Neighbour-Joining trees (Figure 3). The tree based on genotypes are the consensus tree of the 100 different NJ trees from the 100 SNP sets of 44 SNPs and branch support indicated on the tree is derived from the amount of times a specific node is supported by one of the 100 SNP sets (Figure 3A). The distance based analysis of both genotypes and haplotypes reflect the genotype and haplotype STRUCTURE results in that non-Africans are separated from Africans, Bantu-speakers group together and admixed Coloured populations are placed in-between Africans and non-Africans. As with STRUCTURE analysis, distance based analyses based on haplotypes give a better representation of the internal sub-structure

with-in the Khoe-San group. The haplotype based distances reduce the effect of non-African admixture on the Karretjie People and Nama groups and they are grouped with the other Khoe-San groups as seen in the STRUCTURE analysis. The Khwe group in the haplotype analysis is also grouped with the Bantu-speakers rather than the Khoe-San, which also seems a better grouping given their large Bantu-speaking admixture fraction apparent from both genotypic and haplotypic STRUCTURE analyses.

DISCUSSION

This study reports on population structure within the Khoe-San and neighbouring population groups, as revealed by two different types of datasets. For the genotypic datasets, 44 unlinked SNPs were randomly selected from the total 220 SNPs (as explained in the Methods section and Appendix Figure A1). A hundred different such 44 unlinked-SNP datasets were constructed and applied in the calculation of summary statistics, STRUCTURE and distance based analyses. Additionally, a dataset based on short haplotypes was created, consisting of 44 short inferred haplotypes (5 SNPs each) spread over the 22 autosomes. For these 44 short haplotypes we also calculated summary statistics, STRUCTURE and distance based analyses and compared results with results based on genotypic datasets. In addition, the short haplotypes were used to calculate haplotype allele frequencies and private allele frequencies.

Khoe-San groups had the highest frequencies of distinct and private haplotype alleles and non-African groups the lowest, while the Bantu-speakers had intermediate frequencies (Figure 1). Except for the southeast Bantu-speaking group, this pattern is also seen for haplotype heterozygosity estimates. From these estimates it appears if Khoe-San groups have more diversity than Bantu-speaking groups and Africans more

than non-Africans. The difference is not as clear-cut between Bantu-speakers and Khoe-San as between Africans and non-Africans and the large amounts of admixture proportions into and between most of the study groups also will have an effect on estimates. Schlebusch et al., (2012) illustrated the utility of haplotype heterozygosity, haplotype richness and private allelic richness as summary statistics. As seen in the high-density SNP study of Schlebusch et al., (2012), the present study also showed a clear distinction between non-African and African populations, while the pattern within Africa was less clear. A direct comparison between the two studies is not possible since Schlebusch et al., (2012) removed recently admixed individuals from the analysis, which was not done for the present study. However, for both studies it seemed that admixture events (both recent and older events) influence African patterns to a large extent.

Haplotype loci are ideal for estimating the haplotype richness and private haplotype richness summary statistics since it was designed for multi-allelic loci and is not useful for bi-allelic SNP loci. Furthermore, haplotype heterozygosity, rather than SNP heterozygosity, alleviates the effect of ascertainment bias in SNP studies. Even though SNPs from this study have been specifically selected to be heterozygous in African populations, the effect of ascertainment bias can still be seen when comparing haplotype heterozygosity and heterozygosity estimates (Table 1). Firstly, due to general global ascertainment bias of SNPs, the difference between the lowest ranked population (non-African – EUR in both cases) and the highest ranked African population, is more pronounced when considering haplotype heterozygosity, compared to SNP heterozygosity. Secondly, the ascertainment bias introduced by the SNP selection procedure of this specific study (west African Yoruba and African Americans were used as reference populations) is clearly visible. For SNP

heterozygosity, all the Bantu-speaking populations (HER, SEB, DRC), who are genetically close to the west Africans, are among the top populations. When considering haplotype heterozygosity, however, the study-specific ascertainment bias is alleviated and HER and DRC populations move to the lower end of the spectrum, closer to non-African populations. This illustrates the power of haplotypes to alleviate the effect of both general SNP ascertainment bias as well as study-specific SNP ascertainment bias.

STRUCTURE results illustrated different amounts of non-African and Bantu-speaking admixture into the various Khoe-San and Coloured populations. Results supported very low levels of contribution from non-Africans to the northern and central San populations (Ju/'hoansi, !Xun, /Gui,//Gana+Kgalagari, Khwe) and Bantu-speakers. For these populations it is most likely that this low level of non-African cluster contribution is due to shared marker ancestry. Conversely, the southern Khoe-San and Coloured groups (Karretjie People, Nama, Karoo Coloured and Cape Coloured) showed evidence of higher non-African admixture. Schlebusch et al., (2012) also noted the high non-African admixture for these groups, while Pickrell et al., (2012) described high non-African admixture in the Nama (Coloured groups and the Karretjie People were not present in this study). Similar to Schlebusch et al., (2012), the Cape-Coloured group (CAC) had the highest input from the non-African cluster. This is consistent with history, since this group was sampled at Wellington (near Cape Town), which is within the region where the original Cape Colony started. It is well known that during the starting years of the colony very high incidences of mixed unions between colonists, local Khoe and San women and imported slaves occurred (Greeff, 2007, Heese, 1971), which in part gave rise to the Coloured population from this area. Other independent genetic studies also reported high non-

African contributions to Coloured groups from this region (de Wit et al., 2010, Patterson et al., 2010), although gender biased admixture from Khoe-San females were clear (Quintana-Murci et al., 2010). This history is also evident in the STRUCTURE results for another population group in this study; the increased African cluster allocation in the Afrikaner group compared to the European group, illustrate gene-flow from Africans into the Afrikaner population. An increased African cluster is also seen in the South African Indian group (compared to Europeans). The South African Indian population is largely descended from Indians who arrived in South Africa from 1860 onwards as indentured labourers to work on sugarcane plantations (Giliomee and Mbenga, 2010). From the STRUCTURE analysis it appears that the Afrikaner and Indian populations had similar amounts of gene-flow from African populations (Figure 2 and Appendix Table A1).

Evidence of gene-flow from resident Khoe and San groups into the Bantu-speakers, once they expanded into southern Africa, were also observed, since Southern Bantu-speaking groups had higher Khoe-San admixture compared to the Bantu-speakers from the DRC. Schlebusch et al., (2012) and Pickrell et al., (2012) also reported gene-flow from Khoe-San groups into southern Bantu-speakers and Pickrell et al., dated the starting time of admixture to around 1,200 years ago. This date is in agreement with archaeological evidence of when the wave of migrating Bantu-speakers started to arrive in southern Africa (Phillipson, 2005). Excluding the Ju/'hoansi, asymmetric gene-flow between the Bantu-speakers and Khoe-San groups were observed with more gene-flow from the Bantu-speakers into the Khoe-San than vice-versa. The isolated status of the Ju/'hoansi group was confirmed with a far lower contribution from the Bantu-speaking cluster than any of the other Khoe-San groups. Following the Ju/'hoansi, the !Xun group had the highest contribution from the Khoe-

San cluster. The contribution from the Bantu-speaking cluster into the !Xun was more than double that of the Ju/'hoansi group. It is known that the !Xun partly adopted pastoralist practices from surrounding Bantu-speaking groups while the Ju/'hoansi maintained their hunter-gatherer lifestyle, isolating them from pastoralists groups (De Almeida, 1965, Barnard, 1992, Lee, 1979, Marshall, 1960, Guenther, 1986).

An itinerant group from the Karoo region, the Karretjie People, had the third highest contribution from the Khoe-San cluster. This finding supported historical records and local opinion that the Karretjie People are descendant from the /Xam San groups that once lived in the Karoo (Schlebusch et al., 2011, De Jongh, 2002, De Jongh, 2004). Their Coloured neighbours had a much lower input from the Khoe-San cluster.

In support of previous findings based on the classical blood group markers (Jenkins et al., 1971, Jenkins, 1986) and results from the high-density SNP-array studies (Schlebusch et al., 2012, Pickrell et al., 2012), the Khwe had the highest Bantu-speaker admixture of all the Khoe-San groups. Yet, the Khwe showed a much larger contribution from the Khoe-San compared to the Khoe-San component seen in Bantu-speakers, indicating that the Khwe is not merely a Bantu-speaking group that adopted the hunter-gatherer lifestyle and a Khoisan language.

While STRUCTURE results based on genotypic data could not illustrate sub-structure within Khoe-San groups, results from the haplotypic dataset indicated progressively finer level sub-structure within Khoe-San populations as the numbers of assumed clusters were increased. Firstly, the Ju-speakers appeared to cluster together separate from other Khoe-San groups. Furthermore the Karretjie People and Nama groups appeared to cluster together. It is difficult to resolve the relation of the /Gui,//Gana+Kgalagari and the Khwe groups to the other Khoe-San groups. These

two groups both speak languages belonging to the Kalahari-Khoe division of Khoisan, and with initial STRUCTURE clustering they appear to form a common cluster. However, both groups have large amounts of Bantu-speaking admixture which complicates matters and with more allowed clusters, the relationship is not that clear anymore. Also, in the haplotype based NJ tree the /Gui,//Gana+Kgalagari is located on the Khoe-San branch intermediate to the Ju-speakers and the Karretjie+Nama clades, while the Khwe clusters with Bantu-speakers. This might merely be due to higher Bantu-speaking admixture into the Khwe, but in STRUCTURE analyses the Khwe do not form the same homogeneous cluster seen in the /Gui,//Gana+Kgalagari, but instead appears to be comprised of a group of individuals with different genetic ancestries.

Overall analyses based on haplotypes proved to be much more successful in revealing underlying population sub-structure than individual SNPs. Through using short haplotypes we were able to achieve the same level of population distinction as in high-density SNP studies such as described in Schebusch et al., (2012) and Pickrell et al., (2012). The order in which the clusters appeared also corresponds with the high density SNP studies; first the non-African and African populations are differentiated, then the Khoe-San and Bantu-speakers split into different clusters. Thereafter internal Khoe-San structure is revealed with the two Ju-speaking northern San populations (Ju/'hoansi and !Xun) splitting off first, followed by the Khoe speaking Botswana San (/Gui+//Ghana and Khwe) being assigned to their own separate clusters each, next a split between the Ju/'hoansi and !Xun appeared, and thereafter a split between the Karretjie and Nama. It appears thus that by carefully selecting a few SNPs and combining them into haplotypes might be a very efficient and cost-effective way to study population structure.

CONCLUSION

Findings presented here support the division of the study populations into three main groups, Khoe-San, Bantu-speakers and non-African populations. Results thus confirmed the uniqueness of the genetic make-up of the Khoe-San people while illustrating different levels of admixture from the Bantu-speaking and non-African populations into the various San, Khoe and Coloured subgroups. Analyses based on haplotypes indicated further sub-structure within Khoe-San populations and thus demonstrate the effectiveness of combing SNP markers into haplotypes for the inference population structure.

ACKNOWLEDGEMENTS

We are grateful to all subjects who participated in our research and would like to thank members of the Working Group of Indigenous Minorities in Southern Africa (WIMSA) and the South African San Council for facilitating research amongst the San people. We would also like to acknowledge Prof. Michael De Jongh from the University of South Africa (UNISA) who participated in and facilitated the sample collection among the Karretjie People. We furthermore much appreciate the assistance from Prof. Fourie Joubert (Bioinformatics and Computational Biology Unit, University of Pretoria (UP)) for accommodating us in running the STRUCTURE analyses on their cluster computer system at UP. We are very thankful to Prof. Mattias Jakobsson (Uppsala University) who read the draft manuscript and made useful comments. This study was supported by grants awarded to HS from the South African Medical Research Council and the National Research Foundation; HS and CS from the National Health Laboratory Service Research Trust; and CS was

supported by the National Research Foundation (South Africa) and the Wenner-Gren foundation (Sweden).

LITERATURE CITED

Barnard, A. 1992. *Hunters and herders of southern Africa - A comparitive ethnography of the Khoisan peoples.* Cambridge: Cambridge University Press.

Batini, C., G. Ferri, G. Destro-Bisol, et al. 2011. Signatures of the preagricultural peopling processes in sub-Saharan Africa as revealed by the phylogeography of early Y chromosome lineages. *Mol Biol Evol.* 28**:** 2603-13.

Behar, D.M., R. Villems, H. Soodyall, et al. 2008. The dawn of human matrilineal diversity. *Am J Hum Genet.* 82**:** 1130-40.

Bleek, W.H.I. 1862. *A comparative grammar of South African languages. Part I. Phonology.* London.

Browning, S.R. and B.S. Weir. 2010. Population structure with localized haplotype clusters. *Genetics.* 185**:** 1337-44.

Cavalli-Sforza, L.L., P. Menozzi and A. Piazza. 1994. *The History and Geography of Human Genes.* Princeton: Princeton University Press.

Chen, Y.S., A. Olckers, T.G. Schurr, et al. 2000. mtDNA variation in the South African Kung and Khwe-and their genetic relationships to other African populations. *Am J Hum Genet.* 66**:** 1362-83.

Conrad, D.F., M. Jakobsson, G. Coop, et al. 2006. A worldwide survey of haplotype variation and linkage disequilibrium in the human genome. *Nat Genet.* 38**:** 1251-60.

Crawhall, N. 2006. Languages, genetics and archaeology: problems and the possibilties in Africa. In: *The prehistory of Africa* H. Soodyall (ed.) *The prehistory of Africa.* Johannesburg & Cape Town: Jonathan Ball Publishers.

Cruciani, F., R. La Fratta, P. Santolamazza, et al. 2004. Phylogeographic analysis of haplogroup E3b (E-M215). Y chromosomes reveals multiple migratory events within and out of Africa. *Am J Hum Genet.* 74**:** 1014-22.

Cruciani, F., P. Santolamazza, P. Shen, et al. 2002. A back migration from Asia to sub-Saharan Africa is supported by high-resolution analysis of human Y-chromosome haplotypes. *Am J Hum Genet.* 70**:** 1197-214.

De Almeida, A. 1965. *Bushmen and other non-Bantu peoples of Angola.* Johannesburg: Witwatersrand University Press for the Institute for the Study of Man in Africa.

De Jongh, M. 2002. No fixed abode: the poorest of the poor and elusive identities in rural South Africa. *Journal of Southern African Studies.* 28**:** 441-460.

De Jongh, M. 2004. Strangers in Their Own Land: Social Resources and Domestic Fluidity of the Peripatetic Karretjie People of the South African Karoo. In: *Customary Strangers. New perspectives on peripatetic peoples in the Middle East, Africa and Asia* J.C. Berland & A. Rao (eds.) *Customary Strangers. New perspectives on peripatetic peoples in the Middle East, Africa and Asia.* London: Praeger.

De Wit, E., W. Delport, C.E. Rugamika, et al. 2010. Genome-wide analysis of the structure of the South African Coloured Population in the Western Cape. *Human Genetics.* 128**:** 145-153.

Ehret, C. and M. Posnansky. 1982. *The archaeological and linguistic reconstruction of African history.* California: University of California Press.

Falush, D., M. Stephens and J.K. Pritchard. 2003. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics.* 164**:** 1567-87.

Falush, D., M. Stephens and J.K. Pritchard. 2007. Inference of population structure using multilocus genotype data: dominant markers and null alleles. *Mol Ecol Notes.* 7**:** 574-578.

Felsenstein, J. 2004. PHYLIP (Phylogeny Inference Package) version 3.6. *Distributed by the author. Department of Genome Sciences, University of Washington, Seattle.*

Garrigan, D. and M.F. Hammer. 2006. Reconstructing human origins in the genomic era. *Nat Rev Genet.* 7**:** 669-80.

Gattepaille, L.M. and M. Jakobsson. 2012. Combining markers into haplotypes can improve population structure inference. *Genetics.* 190**:** 159-74.

Giliomee, H. and B. Mbenga. 2010. *New History of South Africa.* Cape Town: Tafelberg.

Gordon, R.G. (ed.) (2005) *Ethnologue: Languages of the World,* Dallas, Texas, SIL International.

Greeff, J.M. 2007. Deconstructing Jaco: genetic heritage of an Afrikaner. *Ann Hum Genet.* 71**:** 674-88.

Greenberg, J.H. 1963. *The languages of Africa.* Bloomington, Indiana: Indiana University Press.

Greenberg, J.H. 1972. Linguistic evidence concerning Bantu origins. *J Afr Hist.* 13**:** 189-216.

Guenther, M.G. 1986. Acculturation and assimilation of the Bushmen of Botswana and Namibia. In: *Contemporary Studies on Khoisan* R. Vossen & K.

Keuthmann (eds.) *Contemporary Studies on Khoisan.* Hamburg: Helmut Buske Verlag.

Güldemann, T. 2008. A linguist's view: Khoe-Kwadi speakers as the earliest food-producers of southern Africa. In: *Khoekhoe and the origins of herding in southern Africa* K. Sadr & F.-X. Fauvelle-Aymar (eds.) *Khoekhoe and the origins of herding in southern Africa.* Pietermaritzburg: Southern African Humanities.

Güldemann, T. In Press. Changing profile when encroaching on hunter-gatherer territory: towards a history of the Khoe-Kwadi family in southern Africa. In: *Hunter-gatherers and linguistic history: a global perspective* T. Güldemann, P. Mcconvell & R. Rhodes (eds.) *Hunter-gatherers and linguistic history: a global perspective.* Cambridge: Cambridge University Press.

Heese, J.A. 1971. *Die Herkoms van die Afrikaner, 1657-1867.* Cape Town: A.A. Balkema.

Henn, B.M., C. Gignoux, A.A. Lin, et al. 2008. Y-chromosomal evidence of a pastoralist migration through Tanzania to southern Africa. *Proc Natl Acad Sci U S A.* 105**:** 10693-8.

Huffman, T.N. 1982. Archaeology and the ethnohistory of the African Iron Age. *Ann Rev Anthropol.* 11**:** 133-150.

Jakobsson, M. and N.A. Rosenberg. 2007. CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. *Bioinformatics.* 23**:** 1801-6.

Jakobsson, M., S.W. Scholz, P. Scheet, et al. 2008. Genotype, haplotype and copy-number variation in worldwide human populations. *Nature.* 451**:** 998-1003.

Jenkins, T. 1986. The prehistory of the San and Khoikhoi as recorded in their blood. In: *Contemporary Studies on Khoisan* R. Vossen & K. Keuthmann (eds.) *Contemporary Studies on Khoisan.* Hamburg: Helmut Buske Verlag.

Jenkins, T., H.C. Harpending, H. Gordon, et al. 1971. Red-cell-enzyme polymorphisms in the Khoisan peoples of Southern Africa. *Am J Hum Genet.* 23**:** 513-32.

Jobling, M.A. and C. Tyler-Smith. 2003. The human Y chromosome: an evolutionary marker comes of age. *Nat Rev Genet.* 4**:** 598-612.

Kalinowski, S.T. 2004. Counting Alleles with Rarefaction: Private Alleles and Hierarchical Sampling Designs. *Conservation Genetics.* 5**:** 539-543.

Karafet, T.M., F.L. Mendez, M.B. Meilerman, et al. 2008. New binary polymorphisms reshape and increase resolution of the human Y chromosomal haplogroup tree. *Genome Res.* 18**:** 830-8.

Knight, A., P.A. Underhill, H.M. Mortensen, et al. 2003. African Y chromosome and mtDNA divergence provides insight into the history of click languages. *Curr Biol.* 13**:** 464-73.

Lawson, D.J., G. Hellenthal, S. Myers, et al. 2012. Inference of population structure using dense haplotype data. *PLoS Genet.* 8**:** e1002453.

Le Roux, W. and A. White (eds.) (2004) *Voices of the San,* Cape Town, Kwela Books.

Lee, R.B. 1979. The !Kung San: men, women, and work in a foraging society.). Cambridge: Cambridge University Press.

Li, J.Z., D.M. Absher, H. Tang, et al. 2008. Worldwide human relationships inferred from genome-wide patterns of variation. *Science.* 319**:** 1100-4.

Liu, K. and S.V. Muse. 2005. PowerMarker: an integrated analysis environment for genetic marker analysis. *Bioinformatics.* 21**:** 2128-9.

Marshall, L. 1960. !Kung Bushmen Bands. *Africa.* 30**:** 325-355.

Miller, S.A., D.D. Dykes and H.F. Polesky. 1988. A simple salting out procedure for extracting DNA from human nucleated cells. *Nucleic Acids Res.* 16**:** 1215.

Morin, P.A., K.K. Martien and B.L. Taylor. 2009. Assessing statistical power of SNPs for population structure and conservation studies. *Mol Ecol Resour.* 9**:** 66-73.

Naidoo, T., C.M. Schlebusch, H. Makkan, et al. 2010. Development of a single base extension method to resolve Y chromosome haplogroups in sub-Saharan African populations. *Investig Genet.* 1**:** 6.

Newman, J.L. 1995. *The peopling of Africa.* New Haven, CT: Yale University Press.

Patterson, N., D.C. Petersen, R.E. Van Der Ross, et al. 2010. Genetic structure of a unique admixed population: implications for medical research. *Hum Mol Genet.* 19**:** 411-9.

Phillipson, D. 1993. *African Archaeology.* Cambridge, UK: Cambridge Univ Press.

Phillipson, D. 2005. *African Archaeology.* Cambridge, UK: Cambridge Univ Press.

Pickrell, J.K., N. Patterson, C. Barbieri, et al. 2012. The genetic prehistory of southern Africa. *Nat Commun.* 3**:** 1143.

Pritchard, J.K., M. Stephens and P. Donnelly. 2000. Inference of population structure using multilocus genotype data. *Genetics.* 155**:** 945-59.

Quintana-Murci, L., C. Harmant, H. Quach, et al. 2010. Strong maternal Khoisan contribution to the South African coloured population: a case of gender-biased admixture. *Am J Hum Genet.* 86**:** 611-20.

Rosenberg, N.A. 2002. Distruct: a program for the graphical display of structure results.).

Rosenberg, N.A., S. Mahajan, S. Ramachandran, et al. 2005. Clines, Clusters, and the Effect of Study Design on the Inference of Human Population Structure. *PLoS Genet.* 1**:** e70.

Rosenberg, N.A., J.K. Pritchard, J.L. Weber, et al. 2002. Genetic structure of human populations. *Science.* 298**:** 2381-5.

Schlebusch, C. 2010. Issues raised by use of ethnic-group names in genome study. *Nature.* 464**:** 487; author reply 487.

Schlebusch, C.M., M. De Jongh and H. Soodyall. 2011. Different contributions of ancient mitochondrial and Y-chromosomal lineages in 'Karretjie people' of the Great Karoo in South Africa. *J Hum Genet.* 56**:** 623-30.

Schlebusch, C.M., T. Naidoo and H. Soodyall. 2009. SNaPshot minisequencing to resolve mitochondrial macro-haplogroups found in Africa. *Electrophoresis.* 30**:** 3657-64.

Schlebusch, C.M., P. Skoglund, P. Sjodin, et al. 2012. Genomic Variation in Seven Khoe-San Groups Reveals Adaptation and Complex African History. *Science.* 338**:** 374-379.

Scozzari, R., F. Cruciani, P. Malaspina, et al. 1997. Differential structuring of human populations for homologous X and Y microsatellite loci. *Am J Hum Genet.* 61**:** 719-33.

Scozzari, R., F. Cruciani, P. Santolamazza, et al. 1999. Combined use of biallelic and microsatellite Y-chromosome polymorphisms to infer affinities among African populations. *Am J Hum Genet.* 65**:** 829-46.

Semino, O., A.S. Santachiara-Benerecetti, F. Falaschi, et al. 2002. Ethiopians and Khoisan share the deepest clades of the human Y-chromosome phylogeny. *Am J Hum Genet.* 70**:** 265-8.

Sharp, J. and S. Douglas. 1996. Prisoners of their Reputation? The Veterans of the 'Bushman' Battalions in South Africa. In: *Miscast. Negotiating the Presence of the Bushmen* P. Skotnes (ed.) *Miscast. Negotiating the Presence of the Bushmen.* Cape Town: UCT Press.

Smith, A., C. Malherbe, M. Guenther, et al. 2000. *The Bushmen of Southern Africa.* Cape Town: David Philips Publishers.

Szpiech, Z.A., M. Jakobsson and N.A. Rosenberg. 2008. ADZE: a rarefaction approach for counting alleles private to combinations of populations. *Bioinformatics.* 24**:** 2498–2504.

Tishkoff, S.A., M.K. Gonder, B.M. Henn, et al. 2007. History of click-speaking populations of Africa inferred from mtDNA and Y chromosome genetic variation. *Mol Biol Evol.* 24**:** 2180-95.

Tishkoff, S.A., F.A. Reed, F.R. Friedlaender, et al. 2009. The genetic structure and history of Africans and African Americans. *Science.* 324**:** 1035-44.

Torroni, A., A. Achilli, V. Macaulay, et al. 2006. Harvesting the fruit of the human mtDNA tree. *Trends Genet.* 22**:** 339-45.

Underhill, P.A. and T. Kivisild. 2007. Use of y chromosome and mitochondrial DNA population structure in tracing human migrations. *Annu Rev Genet.* 41**:** 539-64.

Underhill, P.A., G. Passarino, A.A. Lin, et al. 2001. The phylogeography of Y chromosome binary haplotypes and the origins of modern human populations. *Ann Hum Genet.* 65**:** 43-62.

Underhill, P.A., P. Shen, A.A. Lin, et al. 2000. Y chromosome sequence variation and the history of human populations. *Nat Genet.* 26**:** 358-61.

Vansina, J.C. 1990. *Paths in the rainforest. Towards a history of political tradition in equatorial Africa.* London: Currey.

Vigilant, L., M. Stoneking, H. Harpending, et al. 1991. African populations and the evolution of human mitochondrial DNA. *Science.* 253**:** 1503-7.

FIGURE CAPTIONS

Figure 1 - Results from ADZE analyses. A: Mean number of distinct alleles per locus within populations (Y-axis) vs. number of individuals (X-axis). B: Mean number of private alleles per locus within populations (Y-axis) vs. number of individuals (X-axis)

Figure 2 – Individual (A) and population (B) assignments of STRUCTURE runs based on genotypic data (averaged results of STRUCTURE runs of 100 different 44-SNP sets). Cluster assignments of K=2 to K=5 are shown. Individual (C) and population (D) assignments of STRUCTURE runs of the haplotypic dataset. Cluster assignments of K=2 to K=10 are shown.

Figure 3 – Neigbour Joining trees based on distance matrices generated from genotypic datasets (A) and haplotypic dataset (B). The tree based on genotypes (A) is the consensus tree of the 100 different NJ trees from the 100 SNP sets of 44 SNPs and branch support indicated on the tree is derived from the amount of times a specific node is supported by one of the 100 SNP sets. The branch support of the tree based on the haplotypic dataset (B) is the result of 100 bootstrap replicates on the same distance matrix.

APPENDIX

Appendix containing the following additional figures and tables:

Figure A1: SNP selection strategy illustrated on a chromosome. Ten SNPs were chosen for each of the 22 autosomes yielding a total of 220 SNPs.

Table A1: Averaged population cluster assignments (K2-K3) of the STRUCTURE runs from the 100 different 44-SNP sets (Genotypic datasets)

Table A2: SNP panel. A listing of the typed SNP's (rs numbers and base positions) and additional information used in SNP selection.

Table 1 Individuals included in analyses, their group and group-code, place of sampling and place of origin

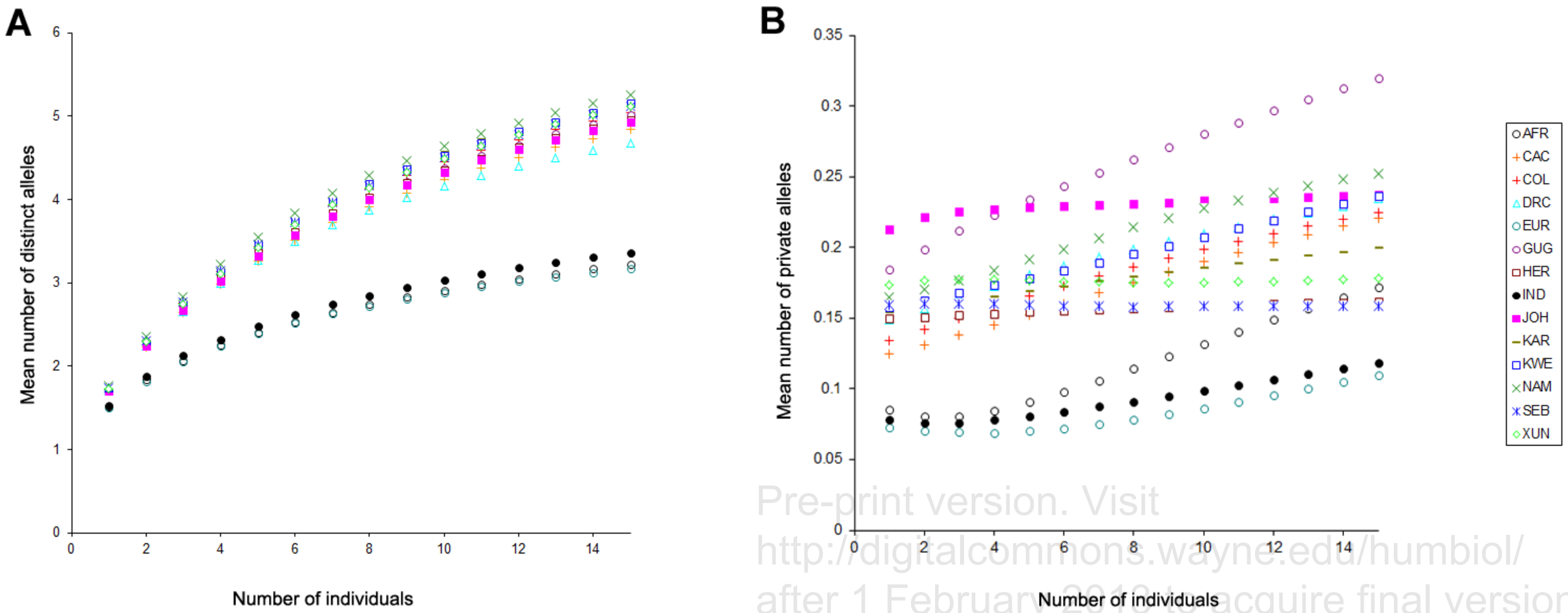
Group name	Group code	Main group	Linguistic classification [1] (Khoisan division shown in descending hierarchy)	Place of sampling (Country) [2]	Place of origin If different from place of sampling	N (publ)[3]	N (new)	N total
Karretjie People [4]	KAR	Coloured (Probable descendants of the /Xam (San))	Probable descendants of: Tuu – !Ui – /Xam	Colesberg (SA)		20	5	25
Karoo Coloured [4]	COL	Coloured	Afrikaans (non-African)	Colesberg (SA)		20	2	22
Cape Coloured	CAC	Coloured	Afrikaans (non-African)	Wellington (SA)		20	0	20
Nama	NAM	Khoe	Khoe – KhoeKhoe – North – Nama-Damara	Windhoek (NM)		20	8	28
/Gui, //Gana and Kgalagari [5]	GUG	Khoe-speaking San	Khoe – Kalahari – West – G//ana – G//ana, G/ui	Kutse Game reserve (BT)		15	6	21
Ju/'hoansi	JOH	San	Ju – Southeast – Ju/'hoan	Tsumkwe (NM)		18	23	41
!Xun	XUN	San	Ju – Northwest – !Xũu	Omega camp (NM) and Schmidtsdrift (SA)	Region surrounding Menongue (AN) [7]	19	26	45
Khwe	KWE	Khoe-speaking San	Khoe – Kalahari – West – Kxoe – Khwe	Omega camp (NM) and Schmidtsdrift (SA)	Caprivi strip and surrounding regions (NM, AN, BT) [7]	17	2	19
Manyanga	DRC	Bantu-speakers	Bantu-speakers (central African)	Luozi (DRC)		0	14	14
Herero	HER	Bantu-speakers	Bantu-speakers (southwestern)	Windhoek (NM)		12	2	14
Southeastern Bantu-speakers [6]	SEB	Bantu-speakers	Bantu-speakers (southeastern)	Various (SA)		20	28	48
Afrikaner	AFR	Non-African	Non-African	Various (SA)		0	15	15
European	EUR	Non-African	Non-African	Various (SA)	Europe and Canada	0	15	15
Indian	IND	Non-African	Non-African	Various (SA)		0	25	25
Total						**181**	**171**	**352**

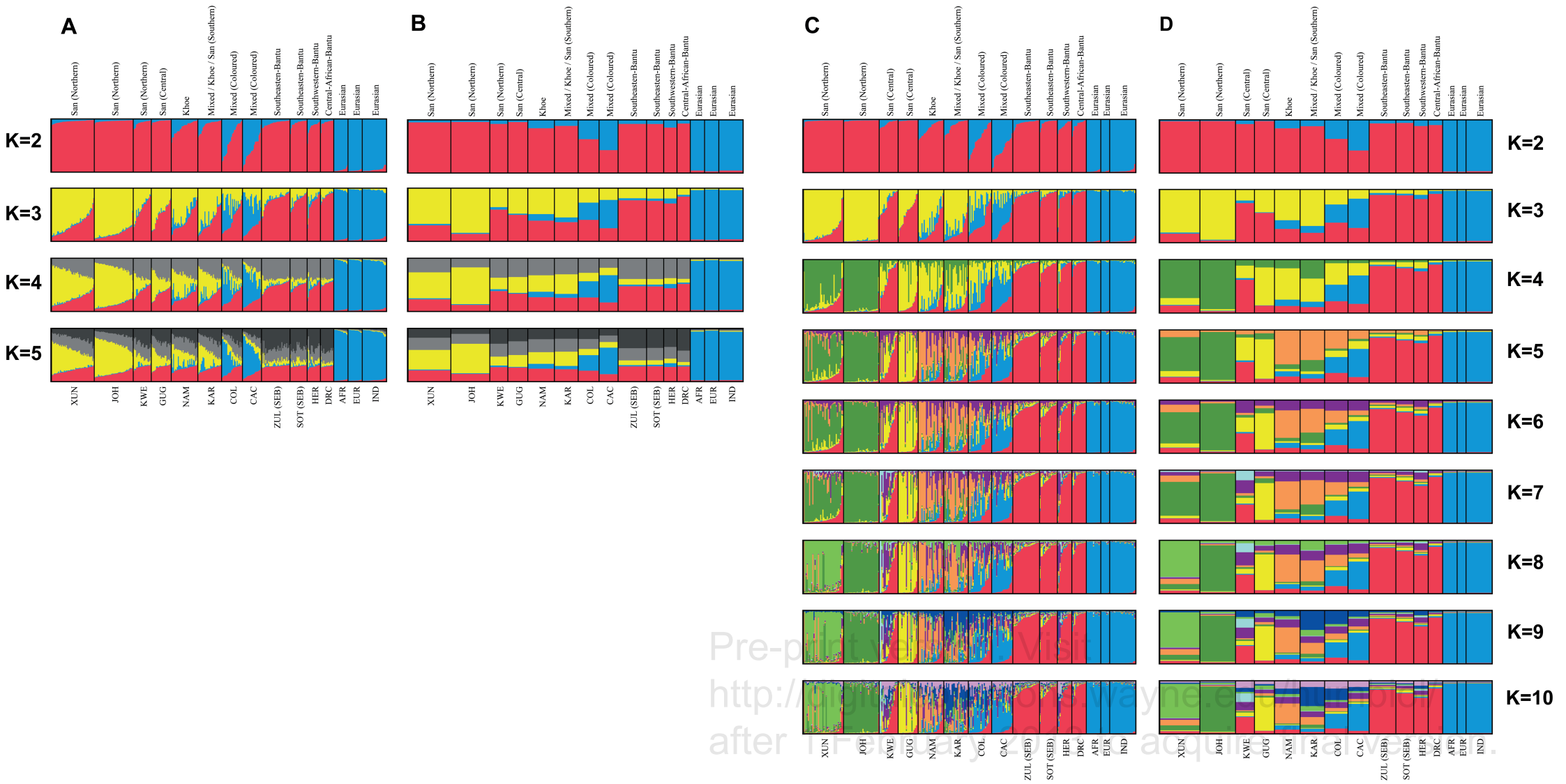
1 Guldemann et al., (2008)
2 Country Abbreviations: AN – Angola, BT – Botswana, DRC – Democratic Republic of Congo, NM – Namibia, SA – South Africa
3 Number of individuals also represented in another published study (Schlebusch et al., 2012).
4 See Schlebusch et al., (2011) for more details on groups
5The GUG group was a mixed group of San and Bantu-speaking individuals who had ancestries from both /Gui and //Gana San groups as well as the Kgalagari Bantu-speaking group
6 Includes Zulu (ZUL), Sotho and Tswana individuals (SOT)
7 According to Sharp and Douglas (1996)

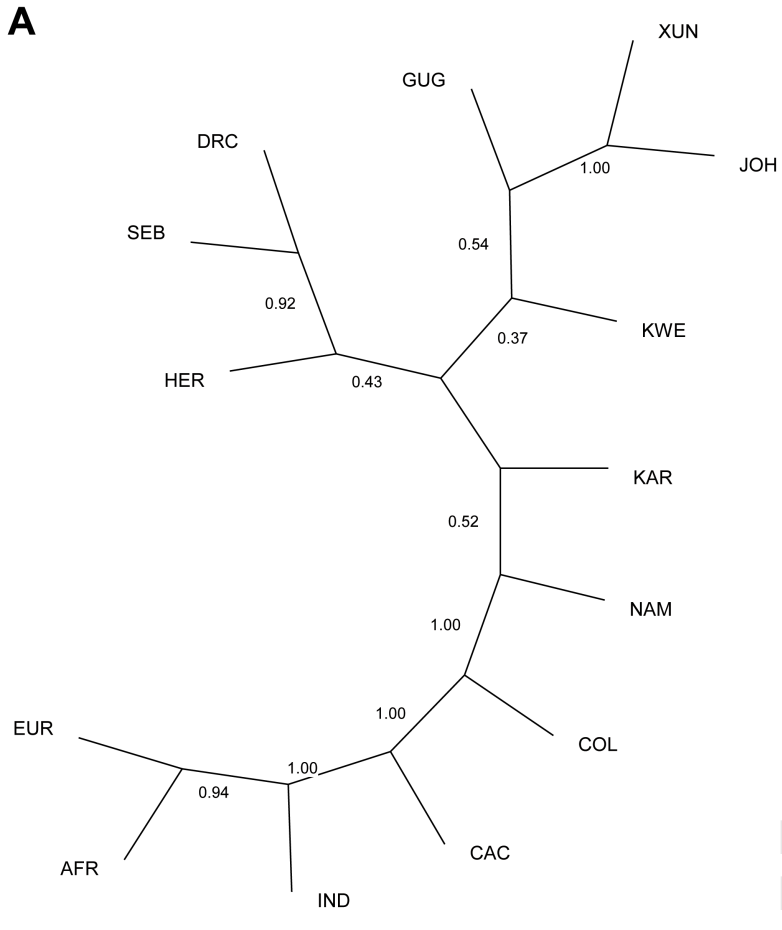
Table 2 Expected Heterozygosity of Haplotypes vs. Genotypes (sorted descending)

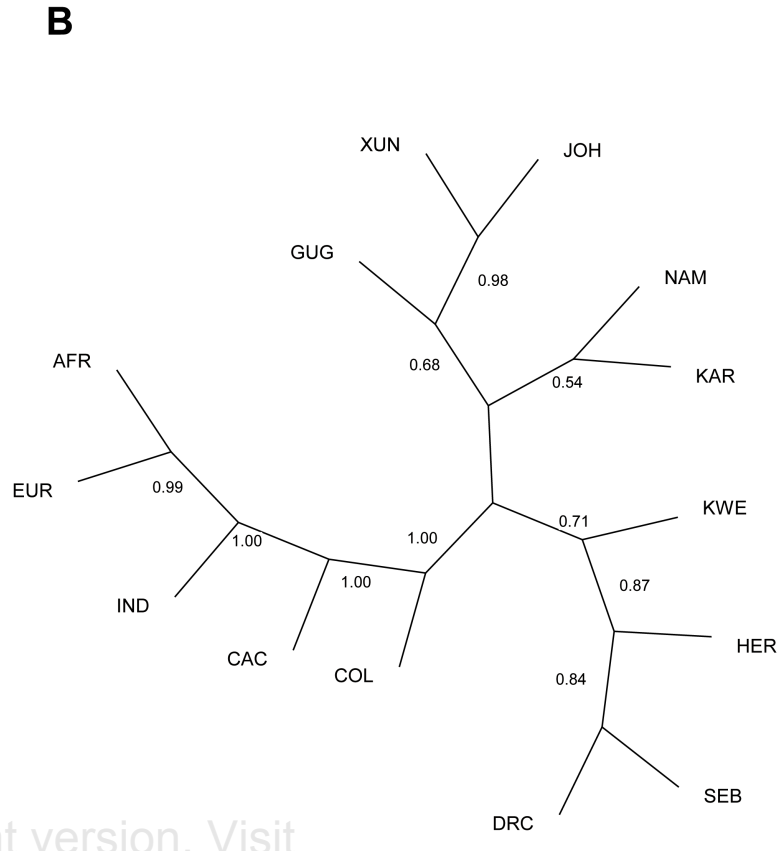
Haplotypes				Genotypes			
Population	**Sample Size**	**Expected Heterozygosity**		**Population**	**Sample Size**	**Mean Expected Heterozygosity**	**STD [1]**
NAM	23	0.742		HER	14	0.424	0.012
SEB	40	0.728		COL	22	0.417	0.014
KAR	22	0.724		SEB	48	0.416	0.014
KWE	17	0.722		DRC	14	0.416	0.013
COL	21	0.722		NAM	28	0.413	0.013
XUN	36	0.716		CAC	20	0.409	0.014
GUG	18	0.702		KWE	19	0.407	0.014
HER	13	0.696		KAR	25	0.397	0.018
JOH	32	0.694		XUN	45	0.384	0.015
DRC	13	0.685		GUG	21	0.383	0.017
CAC	19	0.685		JOH	41	0.361	0.019
IND	23	0.518		IND	25	0.276	0.019
AFR	13	0.490		AFR	15	0.271	0.019
EUR	8	0.468		EUR	15	0.253	0.020

1 Standard deviation between the expected heterozygosity of 100 sets of 44 unlinked SNPs









Appendix

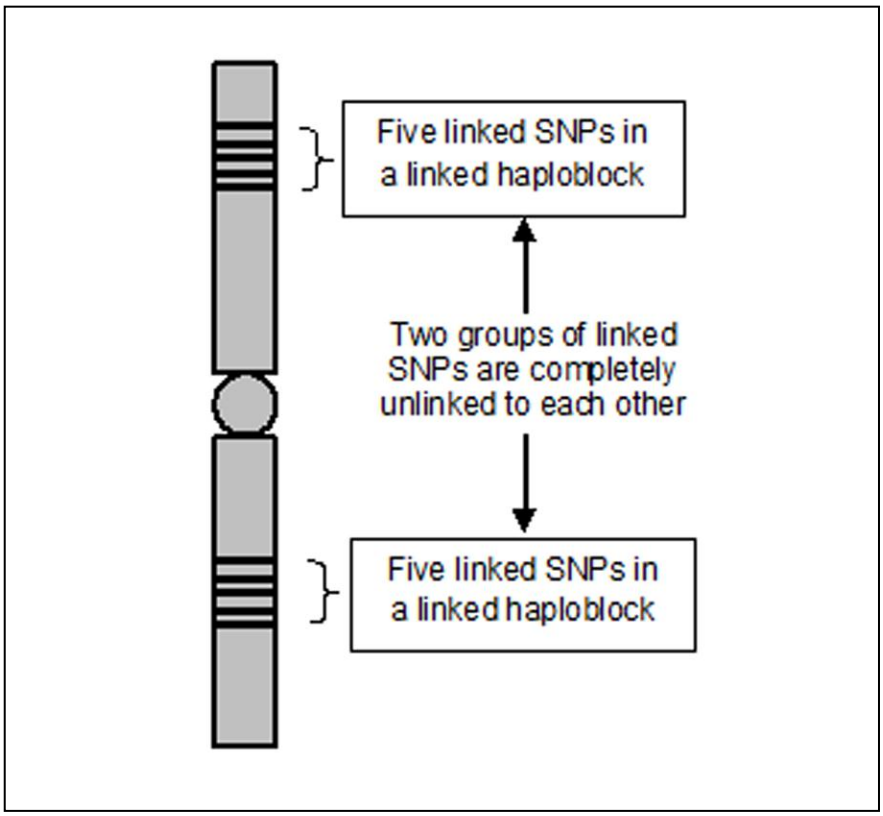


Figure A1. SNP selection strategy illustrated on a chromosome. Ten SNPs were chosen for each of the 22 autosomes yielding a total of 220 SNPs.

Table A1: Averaged population cluster assignments (K2-K3) of the STRUCTURE runs from the 100 different 44-SNP sets (Genotypic datasets)

K	Pop	Cluster 1	Cluster 2	Cluster 3
2	XUN	0.038	0.962	
2	JOH	0.03	0.97	
2	KWE	0.067	0.933	
2	GUG	0.032	0.969	
2	NAM	0.149	0.851	
2	KAR	0.106	0.894	
2	COL	0.361	0.639	
2	CAC	0.571	0.429	
2	SEB	0.068	0.932	
2	HER	0.138	0.862	
2	DRC	0.06	0.94	
2	AFR	0.964	0.036	
2	EUR	0.979	0.021	
2	IND	0.963	0.037	
3	XUN	0.029	0.674	0.296
3	JOH	0.022	0.846	0.132
3	KWE	0.048	0.353	0.598
3	GUG	0.023	0.478	0.498
3	NAM	0.127	0.487	0.386
3	KAR	0.091	0.551	0.358
3	COL	0.332	0.267	0.402
3	CAC	0.543	0.216	0.242
3	SEB	0.042	0.185	0.773
3	HER	0.097	0.187	0.716
3	DRC	0.035	0.117	0.848
3	AFR	0.949	0.027	0.024
3	EUR	0.968	0.016	0.017
3	IND	0.946	0.026	0.028

Table A2: SNP panel. A listing of the typed SNP's (rs numbers and base positions) and additional information used in SNP selection.

Chromosome	Group on chromosome	SNP ID (rs)	Base Position	Distance from previous marker	Yoruba MAF
1	1	rs7523071	185581438		32
1	***1***	***rs1445667***	***185582609***	***1171***	***32***
1	1	rs1445670	185587536	4927	42
1	1	rs6660605	185594475	6939	32
1	1	rs6666285	185594691	216	44
1	2	rs6702432	243839090		26
1	2	rs7366424	243845655	6565	27
1	2	rs7555211	243848391	2736	36
1	2	rs1954187	243851004	2613	9
1	***2***	***rs10399826***	***243861576***	***10572***	***38***
2	1	rs2373901	40769550		50
2	1	rs882007	40781807	12257	20
2	1	rs6755751	40782253	446	47
2	1	rs3851315	40785189	2936	37
2	1	rs11124754	40787764	2575	32
2	2	rs6743609	78370721		27
2	2	rs6715934	78379067	8346	23
2	2	rs2839828	78382969	3902	39
2	2	rs1837144	78383601	632	50
2	2	rs1816652	78388857	5256	17
3	1	rs1987888	4053654		24
3	1	rs1087817	4063576	9922	33
3	1	rs317575	4063809	233	N/A
3	1	rs317530	4069293	5484	34
3	1	rs317534	4074043	4750	49
3	2	rs4624549	189144204		48
3	2	rs2590451	189147479	3275	42
3	2	rs567713	189151423	3944	47
3	2	rs2679506	189154725	3302	28
3	2	rs522833	189160082	5357	27
4	1	rs9998475	13325188		26
4	1	rs1352786	13326354	1166	26
4	1	rs1948354	13334081	7727	26
4	1	rs6837122	13335534	1453	24
4	1	rs1032358	13338502	2968	19
4	2	rs10084822	172054953		29

4	2	rs9312493	172061519	6566	31
4	2	rs10004230	172066255	4736	17
4	2	rs1403213	172075840	9585	43
4	***2***	***rs10002204***	***172096780***	***20940***	***21***
5	1	rs1366370	66593667		39
5	1	rs755877	66593979	312	45
5	1	rs1593948	66594316	337	47
5	1	rs7715561	66598715	4399	37
5	1	rs919308	66604140	5425	17
5	***2***	***rs165073***	***163963822***		***31***
5	2	rs1363174	163978188	14366	N/A
5	2	rs250597	163980289	2101	30
5	2	rs10515884	163985604	5315	41
5	2	rs1421905	163990354	4750	38
6	1	rs9505359	809219		22
6	1	rs884126	815244	6025	27
6	1	rs885450	815563	319	N/A
6	1	rs873560	820559	4996	24
6	1	rs6916756	825467	4908	23
6	2	rs6912046	79193277		45
6	2	rs2223722	79197714	4437	46
6	2	rs926654	79202638	4924	36
6	2	rs9361404	79205477	2839	21
6	2	rs9448411	79208314	2837	32
7	1	rs2592859	35206935		31
7	1	rs731015	35212110	5175	25
7	1	rs2541911	35216715	4605	37
7	1	rs2250212	35221258	4543	7
7	1	rs2592848	35230892	9634	22
7	2	rs7806350	144859843		49
7	2	rs1523729	144867554	7711	27
7	2	rs2888245	144871885	4331	24
7	2	rs1523723	144877013	5128	20
7	***2***	***rs6954212***	***144880096***	***3083***	***28***
8	1	rs871565	18152103		39
8	1	rs1493029	18165651	13548	29
8	1	rs902960	18168085	2434	38
8	***1***	***rs7846103***	***18170309***	***2224***	***22***
8	1	rs2131422	18178912	8603	23
8	2	rs2385226	126751178		17
8	2	rs4871628	126752121	943	27

8	2	rs7838054	126753324	1203	27
8	2	rs1159478	126757397	4073	N/A
8	2	rs7460157	126761038	3641	22
9	1	rs10966574	24919668		42
9	1	rs7025715	24924491	4823	37
9	1	rs7871011	24925087	596	47
9	1	rs4085752	24931125	6038	14
9	1	rs1461333	24936349	5224	42
9	2	rs1927239	123675437		21
9	2	rs2489161	123678034	2597	28
9	2	rs562239	123679804	1770	21
9	2	rs4836945	123689332	9528	21
9	2	rs2768818	123690135	803	28
10	1	rs9663972	60527538		20
10	1	rs6481457	60531364	3826	42
10	1	rs733341	60533393	2029	46
10	1	rs11006373	60539023	5630	45
10	1	rs7921026	60541895	2872	27
10	2	rs7094944	109799612		37
10	2	rs10509859	109803462	3850	23
10	2	rs1125798	109808286	4824	25
10	2	rs7073564	109813235	4949	23
10	2	rs1556592	109819760	6525	35
11	1	rs7124156	13198502		42
11	1	rs900141	13204100	5598	20
11	1	rs900142	13204831	731	22
11	1	rs7117211	13205223	392	32
11	1	rs7107711	13212114	6891	43
11	2	rs2042599	127235817		34
11	2	rs1812931	127240375	4558	30
11	2	rs1364777	127242208	1833	27
11	2	rs1107869	127249002	6794	27
11	2	rs10893778	127253038	4036	27
12	1	rs917589	3412660		17
12	1	rs917587	3412936	276	16
12	1	rs2878578	3413587	651	47
12	1	rs6489468	3421275	7688	34
12	1	rs7961141	3424976	3701	45
12	2	rs855228	101400231		29
12	2	rs855224	101405390	5159	35
12	2	rs855218	101409109	3719	35

12	2	rs855211	101413277	4168	32
12	2	rs35746	101417107	3830	47
13	1	rs4769191	21547069		35
13	1	rs1323170	21547219	150	19
13	1	rs4770238	21548179	960	37
13	1	rs9316743	21548512	333	45
13	1	rs1323172	21550247	1735	44
13	2	rs978089	85554112		21
13	***2***	***rs4910994***	***85559270***	***5158***	***41***
13	2	rs1029143	85563006	3736	41
13	2	rs9594117	85578891	15885	20
13	2	rs1413441	85580898	2007	19
14	1	rs2383584	33849679		21
14	1	rs7143582	33852799	3120	33
14	1	rs1958572	33858595	5796	47
14	1	rs1958574	33867066	8471	15
14	1	rs1958579	33870654	3588	13
14	2	rs1241743	91751928		40
14	2	rs1241745	91752315	387	36
14	2	rs1956413	91753943	1628	44
14	2	rs1956414	91758924	4981	40
14	2	rs1741443	91774327	15403	47
15	1	rs722150	31201795		N/A
15	1	rs4780082	31202774	979	23
15	1	rs1988447	31204618	1844	14
15	1	rs7181962	31204650	32	46
15	1	rs8023846	31211066	6416	35
15	2	rs920921	66573339		41
15	2	rs1373697	66577067	3728	35
15	2	rs895133	66580703	3636	34
15	2	rs2084032	66582870	2167	37
15	2	rs895131	66583554	684	27
16	1	rs1848824	61630443		44
16	1	rs153322	61631942	1499	50
16	1	rs153341	61644707	12765	23
16	1	rs1605960	61655814	11107	27
16	1	rs198007	61678146	22332	28
16	2	rs1510205	84851316		17
16	2	rs2883250	84859632	8316	22
16	2	rs2696815	84859844	212	40
16	2	rs717482	84862498	2654	16

16	2	rs1027910	84866445	3947	17
17	1	rs2007643	52084308		38
17	1	rs6503752	52088714	4406	38
17	1	rs714832	52093256	4542	29
17	1	rs10491158	52099116	5860	17
17	1	rs1019117	52103128	4012	27
17	2	rs7222022	66763060		38
17	2	rs2158906	66769428	6368	21
17	2	rs724856	66776439	7011	N/A
17	2	rs2190461	66787482	11043	48
17	2	rs6501466	66789877	2395	36
18	1	rs2940757	34847593		32
18	1	rs2958610	34848055	462	45
18	1	rs1509219	34852830	4775	32
18	1	rs9304198	34854813	1983	17
18	1	rs8083419	34856469	1656	47
18	2	rs165130	73464384		37
18	2	rs905443	73464575	191	40
18	2	rs165128	73464782	207	18
18	2	rs9952646	73470415	5633	37
18	2	rs2407139	73472582	2167	21
19	1	rs7256520	36812013		42
19	1	rs8100570	36814702	2689	42
19	1	rs892210	36817849	3147	N/A
19	1	rs8112540	36818052	203	31
19	1	rs8101359	36822617	4565	33
19	2	rs1654338	43228193		24
19	2	rs734204	43231828	3635	25
19	2	rs941037	43235466	3638	25
19	2	rs1725467	43235743	277	25
19	2	rs1725504	43238719	2976	35
20	1	rs6085916	7112725		21
20	1	rs1033604	7126839	14114	22
20	1	rs1016264	7128675	1836	N/A
20	1	rs6133401	7129330	655	22
20	1	rs6117693	7135874	6544	22
20	2	rs2424383	21514717		12
20	2	rs1014889	21518837	4120	31
20	2	rs1014890	21519183	346	N/A
20	2	rs1074606	21521722	2539	30
20	2	rs6035902	21530338	8616	36

21	1	rs150210	18284183		37
21	1	rs197562	18285788	1605	26
21	1	rs2824593	18290483	4695	22
21	1	rs158077	18294083	3600	49
21	1	rs1505265	18296572	2489	40
21	2	rs8131079	24467571		37
21	2	rs7280999	24469539	1968	37
21	2	rs1024318	24475083	5544	22
21	2	rs1910605	24480779	5696	49
21	2	rs1910622	24483502	2723	49
22	1	rs137462	31940397		N/A
22	1	rs9306274	31940642	245	32
22	1	rs137472	31945136	4494	32
22	1	rs118033	31945610	474	18
22	1	rs137475	31951775	6165	27
22	2	rs2413378	34796843		17
22	2	rs715550	34797853	1010	25
22	2	rs715546	34798045	192	24
22	2	rs7286844	34804171	6126	20
22	2	rs739203	34805381	1210	42
			AVE	**4347**	
			STD	**3730.8**	
			Min	**192**	
			Max	**22332**	

SNPs in ***Bold Italic*** were excluded due to poor quality